SSLA, **29**, 449–484. Printed in the United States of America.
DOI: 10.1017/S0272263107070271

SECOND LANGUAGE IDIOM LEARNING IN A PAIRED-ASSOCIATE PARADIGM

Effects of Direction of Learning, Direction of Testing, Idiom Imageability, and Idiom Transparency

Margarita P. Steinel and Jan H. Hulstijn
University of Amsterdam

Wolfgang Steinel
Leiden University

In a paired-associate learning (PAL) task, Dutch university students (n = 129) learned 20 English second language (L2) idioms either receptively or productively (i.e., L2-first language [L1] or L1-L2) and were tested in two directions (i.e., recognition or production) immediately after learning and 3 weeks later. Receptive and productive performance was affected by direction of learning. This finding parallels findings from PAL experiments on L2 individual-word learning. On a productive test, productive learners had a sizable advantage over receptive learners, whereas on recognition, receptive learners outperformed productive learners. Two idiom characteristics, imageability (capacity to evoke a mental image) and transparency (overlap between literal and figurative meaning), as assessed in a norming

This study is based on the first author's MA research. We would like to thank the four anonymous *SSLA* reviewers for their helpful comments on this manuscript. Any errors or omissions remain our own.

Address correspondence to: Jan H. Hulstijn, Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands; e-mail: j.h.hulstijn@uva.nl; or Margarita P. Steinel, Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 134, 1012 VB Amsterdam, The Netherlands; e-mail: m.p.steinel@uva.nl.

study by an independent sample (n = 80), qualified these findings. Indicating the importance of dual coding in idiom learning, imageability predicted performance, and receptive learning was particularly inefficient for low imageable idioms. Transparency was a weaker predictor of performance and only affected recognition.

Much of lexis consists of sequences of words that have a strong tendency to occur together in discourse, including a wide and motley range of expressions such as phrasal verbs, compounds, idioms, and collocations (referred to collectively as multiword lexical items, prefabricated units, prefabs, phraseological units, fixed phrases, formulaic sequences, etc.). Phraseological performance is, by general consent, an important component of second language (L2) fluency, yet nonnative speakers' speech and writing display a significantly lower collocational density (Howarth, 1998) and their knowledge of complex lexical units in their L2 is limited compared to that of native speakers (Arnaud & Savignon, 1997; Moon, 1997). L2 learners' knowledge of L2 multiword units such as idioms and collocations is not on a par with their knowledge of L2 vocabulary in general. Clearly, this constitutes a major difficulty for L2 learners. The investigation of which conditions facilitate learning and, consequently, comprehension and production will help us find out more about the locus of this difficulty. In the setting of a paired-associate learning (PAL) experiment, we investigate, in particular, the effects of direction of learning, direction of testing, imageability, and transparency of L2 idioms on immediate and long-term retention. We will first situate our study in the literature on PAL, imageability, transparency, and long-term retention.

PAIRED-ASSOCIATE LEARNING

Paired-associate learning has been the subject of a copious amount of empirical research, in particular in the 1950s and the 1960s (e.g., Crothers & Suppes, 1967; Underwood & Schulz, 1960). In early studies that applied the PAL paradigm, participants learned to pair a familiar first language (L1) word (the response) to another familiar L1 word (the stimulus). The aim of these studies was to investigate the establishment of within-language associative connections. As Griffin and Harley (1996) summarized, evidence from earlier PAL studies is inconclusive. The bone of contention seems to be the question of which element of the paired associate has a more important role to play in the establishment of a memory association between the two. Some studies stressed the importance of the element in the response position, others did not find any differences, and yet others claimed that the element in the stimulus position was the more important one.

The PAL paradigm has also been used in studies on L2 vocabulary learning in which participants paired an unfamiliar L2 word with a familiar L1 word. Both the learning and the testing task can be either receptive or productive, with the L2 word in the stimulus or response position, respectively. The results of a study conducted by Schuyten (1906) with Dutch pupils learning French, German, and English indicated that (a) receptive retention was always substantially higher than productive retention and (b) receptive learning led to a substantial amount of productive knowledge and vice versa. The results of a study conducted by Stoddard (1929), who asked American high school students without any prior knowledge of French to learn French words, suggested that (a) receptive performance was significantly higher than productive performance; (b) the best results on the receptive test were obtained when participants learned receptively, and, similarly, the best results on the productive test were obtained when participants learned productively; and (c) productive learning led to a considerable amount of receptive knowledge and vice versa.

More recently, Griffin and Harley (1996), Schneider, Healy, and Bourne (2002), and Mondria and Wiersma (2004) revived the full PAL paradigm by investigating the effects of direction of learning and direction of testing in L2 vocabulary learning. Griffin and Harley asked British comprehensive school students aged 11–13, who had had 6 months of formal instruction in French, to learn 20 word pairs (either L1-L2 or L2-L1) in 8 min and tested them either in the same direction as in the learning session or in the opposite direction. Testing took place immediately after learning and, without prior announcement, 3, 7, and 28 days after learning. From the results of their study, Griffin and Harley concluded that the L1-L2 direction of learning is "the better all-purpose direction, more effective than the L2-L1 for the more difficult production task" (p. 454). With regard to retention over time, direction of learning was not found to influence the strength of the association over time.

Schneider et al. (2002) manipulated translation direction within a paired-associate task framework in order to explore rates of retention and transfer. In their first experiment, 25 cue-target vocabulary pairs each involving a French word and its translation were presented to American college students without prior knowledge of French. Retention and transfer were measured in two sessions. In the first session, the direction of learning for half of the participants was L2-L1, whereas the direction of learning for the other half was L1-L2. In all cases, the direction of learning and the direction of testing at the immediate test at the end of the first session were the same; that is, in the immediate test session, Schneider et al. did not collect data on what Griffin and Harley (1996) would call backward association. In the second session, one half of each of the two groups was tested and retrained in the same direction as during the first session and the other half was tested and retrained in the reverse direction.

One of the main claims of the Schneider et al. (2002) study was that participants trained in the context of the more difficult task (L1-L2) had, despite infe-

rior initial performance, an advantage on the delayed test, especially when they had to do the more difficult L1-L2 test. This interpretation of the results supports the general assumption that learning tasks under more difficult conditions (in this case L1 cues and L2 responses) yields inferior learning and immediate retention but less loss across retention intervals than learning tasks under easier conditions.

Schneider et al.'s (2002) second experiment had a similar design. One of the few differences was that participants were pretrained on the orthography of half of the L2 words. In accordance with their previous assumptions and findings, Schneider et al. found that pretraining by decreasing task difficulty enhanced initial learning but not retention and transfer. Again, initial performance was better when participants were trained with L2 cues, but this did not lead to better delayed retention and transfer. In summary, Schneider et al.'s main conclusion was that greater difficulty of the learning task decreased initial performance but led to better delayed performance when the difficulty was manipulated by translation direction.

Mondria and Wiersma (2004) studied the effect of the combination of receptive and productive learning versus receptive learning alone or productive learning alone on receptive and productive retention, respectively. According to the combination hypothesis, learning both receptively and productively should lead to better and more stable receptive retention performance than receptive learning alone. Dutch secondary school students learned 16 French-Dutch pairs of words (without context) in one of three learning conditions (receptive, productive, or both) and were tested twice (immediately and about 2 weeks later) in either direction of testing (receptive vs. productive). On the delayed receptive retention test, no significant difference was found between the receptive and the receptive plus productive learning condition. Similarly, on the delayed productive retention test, those who learned both receptively and productively did not differ significantly from those who only learned productively. However, against the authors' expectations, productive learning alone led to significantly better performance than the combination of receptive and productive learning on the immediate productive test. With regard to receptive retention, participants who had learned receptively did better than participants who had learned productively. Finally, concerning productive retention, productive learning led to significantly better results than receptive learning.

Mondria and Wiersma (2004) also explored an issue that all other related studies have shied away from discussing in detail—the issue of the degree of difficulty inherent in the direction of learning (productive vs. receptive). The authors suggested that relevant studies provide evidence for the relative difficulty of productive learning compared to receptive learning. This, they claimed, is "evidenced by the fact that in all the experiments the mean scores on the productive retention tests were lower than those on the receptive retention tests" (Mondria & Wiersma, p. 86). By comparing receptive and productive retention resulting from the combination of receptive and productive

learning, Mondria and Wiersma found that receptive retention was significantly higher, and they concluded that productive learning (in terms of retention) is indeed more difficult. Two explanations are put forth for this finding. The "amount of knowledge" explanation (Nation, 2001, p. 28) suggests that the greater difficulty of productive retention is due to the necessity of having more precise knowledge of word form in order to use a word in a productive way. The so-called "access explanation" (Nation, p. 29) is based on the idea that a new L2 word is only (receptively) linked to its L1 equivalent and not to other L2 words in the lexical system, unlike the L1 equivalent, which has a host of links to other L1 words, all of which can be conceptualized as different competing paths along which the L1 word can be accessed.

Taking as a starting point the idea that the best performance can be achieved when the direction of learning is the same as the direction of testing, Mondria and Wiersma (2004) explored the question of whether the effect of type of test (productive vs. receptive) overrides the effect of correspondence between type of learning and type of test. By comparing receptive and productive delayed retention resulting from productive learning, they found that receptive retention was higher, which they interpreted as evidence that the effect of type of test is greater than the effect of correspondence between type of learning and type of test. On the immediate test, however, the opposite pattern emerged: Productive learning led to significantly better productive than receptive retention, which suggests that on retention, the effect of correspondence between type of learning and type of test is greater than the effect of type of test.

A practical conclusion suggested in the study is that when productive knowledge is the measure and goal of L2 vocabulary learning, learning words productively is the more effective approach. Griffin and Harley (1996) also concluded that learning L2 vocabulary productively is the more versatile direction for the demands of both receptive and productive performance.

One alternative perspective on the question of how direction of learning prepares learners for performance on different tests could be provided by the notion of transfer-appropriate processing (TAP), which accounts for retention performance in terms of the (in)compatibility between learning and retention tasks (Bransford, Franks, Morris, & Stein, 1979; see also Hulstijn, 2003). In essence, the TAP principle suggests that the type of initial processing of specific stimuli will facilitate subsequent processing of the same stimuli, depending on the extent to which the two overlap; that is, productive learning, which can be assumed to share more common processes with productive testing than receptive learning does, is likely to meet the requirements of the more difficult test more efficiently because of the greater similarity between the two tasks.

Thus, it seems sound to assume that the general principles and mechanisms discussed here as possible explanations of the findings with regard to L2 individual-word learning will apply to the learning of multiword units as well. Learning idioms in the L1-L2 direction, being more difficult, might lead

to more elaborate processing and more remembering. An implication of the TAP principle would be that receptive learning, on the one hand, will possibly be adequate for the demands of the receptive test but less so when productive knowledge is tested. Productive learning, on the other hand, will probably be much less disadvantaged by the more difficult productive test because this would correspond to the direction in which knowledge was initially acquired. We therefore propose two hypotheses related to the direction of leaning and testing:

(1) Directionality hypotheses
   a. Performance on a productive test (i.e., L1-L2) will be better when idioms have been learned productively (L1-L2) rather than receptively (L2-L1).
   b. On a receptive test (i.e., L2-L1), the performance of L1-L2 learners will not be superior to the performance of L2-L1 learners.

IMAGEABILITY

In this section, we look at some studies that have examined imageability as a determinant of learning. Paivio, Yuille, and Smythe (1966) looked at the effects of an image-evoking potential or imagery of words on recall in monolingual PAL experiments. Their results suggested that imagery of both the stimulus and the response were found to be good predictors of learning (e.g., *shoe-idea* and *idea-shoe* were learned better than *idea-truth* or *truth-idea*). The effect of imagery on learning was explained as a two-code representational mechanism in which information can be encoded as visual images, verbal representations, or both. The information required for recall can then be retrieved from either of the two codes or from both. The nonverbal code is conceived of as "a supplementary coding system, which enhances the probability that items will be correctly retrieved on test trial" (Paivio, 1969, p. 257). This is the gist of the so-called dual-coding theory.

In addition to notions of concreteness and meaningfulness, as proposed by Paivio et al. (1966), similar constructs have been proposed, such as ease of predication (Jones, 1985) and context availability (Schwanenflugel, Harnishfeger, & Stowe, 1988). One common feature of these different theories is that things that are grounded in perceptual experience and are analyzed visually or perceptually are imageable and meaningful. Ellis and Beaton (1993) suggested that imageability effects in verbal learning reflect the fact that visual imageability confers meaning (cf. Lakoff & Johnson, 1980). In word learning it can indeed be difficult to distinguish precisely between concreteness (the availability of direct sensory referents for a concept) and imageability. De Groot and Keijzer (2000) and De Groot and Poot (1997), for example, used the two terms interchangeably. The distinction is probably easier to uphold with regard to idioms, in that concreteness would be a relevant characteristic at the level of constituent words, whereas imageability would provide a particularly good perspective to the idiom as a whole. As the main emphasis of this study is

idiom learning rather than word learning, imageability is a more relevant characteristic than concreteness.

There is an ongoing debate on the issue of how lexical and conceptual representations are related in the bilingual lexicon (see Kroll & Tokowicz, 2005, for a review). In the revised hierarchical model (Kroll & Stewart, 1994), the links between L1 and L2 words and conceptual memory differ in strength in such a way that the link between L1 and conceptual memory is stronger than the link between L2 and conceptual memory. An implication of this model for studies investigating the role of imageability of words in bilingual experiments might be that the effect of imageability is greater when the L1 words are given and the L2 words have to be produced (i.e., when productive retention is needed) than when the L2 words function as stimuli (i.e., when receptive retention is tested).

De Groot and colleagues (De Groot, 2006; De Groot & Hoeks, 1995; De Groot & Keijzer, 2000; De Groot & Poot, 1997) investigated the role of imageability on learning and translation performance. Overall, their results demonstrated that words with higher imageability were learned better and facilitated translation more than words with lower imageability.

In a study on word learning using the keyword method, Ellis and Beaton (1993) found that keyword imageability is more important when translating from L2 to L1 than from L1 to L2. As the keyword belongs more clearly to the stimulus side of the association when translating from L2 to L1, this finding parallels one recurrent conclusion of PAL experiments: The imageability of the stimulus is more important than that of the response (e.g., Paivio, 1971).

The role of images evoked by L2 idioms was discussed by Boers, Demecheleer, and Eyckmans (2004). They conducted two experiments that investigated whether the learning strategy of etymological elaboration (EE)—raising learners' awareness of the literal origins or source domains of figurative expressions—is equally effective for the retention of etymologically opaque idioms as for the retention of etymologically transparent ones. "Barking up the wrong tree," for instance, can be traced back to the domain of hunting (in a foxhunt, the dogs sometimes corner the fox in a tree and bark up at that tree) and would be an example of an etymologically transparent idiom. The results revealed that EE could be successfully applied to both types of idioms. The beneficial mnemonic effect of EE can be explained with reference to dual-coding theory: Awareness of the origins or a source domain of an idiom might help form a mental image of the specific context that motivated the first occurrences of that idiom. Storing verbal information as a mental image provides an extra pathway for recall, as information is encoded in a dual fashion. Also, EE is likely to help learners realize that some idioms are motivated, as the figurative meaning becomes more easily explicable in the light of information about the etymological origin of those idioms. Bortfeld (2002) suggested that the analysis of idiom surface form that takes place while forming a mental image of an idiom might be beneficial to L2 learners in enabling them to understand the mapping between surface forms and conceptual structures.

The facilitating effect of imageability, combined with our assumption that the component that can benefit more from dual coding is actually the initially unfamiliar L2 expression, might lead to a greater facilitating effect of imageability in the L2-L1 learning condition. As such, we propose three hypotheses related to imageability in L2 idiom learning:

(2) Imageability hypotheses
   a. In each of the two test conditions (production and comprehension), performance on idioms that are high on imageability will be better than performance on idioms that are low on imageability.
   b. This effect will be stronger for idioms that have been learned in the L2-L1 learning direction.
   c. In each of the two test conditions, high positive correlations will be observed between test performance and the (rating of one's) ability to visualize the idioms to be learned.

TRANSPARENCY

Geeraerts (1995) distinguished syntagmatic transparency, or isomorphism, from motivation, or paradigmatic transparency. Syntagmatic transparency is the

> one-to-one correspondence between the formal structure of the expression and the structure of its semantic interpretation, in the sense that there exists a systematic correlation between parts of the semantic value of the expression as a whole and the constituent parts of that expression. (Geeraerts, p. 61)

An example would be "to take the bull by the horns," which is isomorphic because there is a one-to-one mapping between the meanings of the constituent parts of the idiom and the elements of the global idiomatic meaning, which can be paraphrased as "to tackle a problem or a difficulty at the central, most dangerous, or most difficult point." For a classification of idiomatic expressions based on syntagmatic transparency, see Cacciari and Glucksberg (1991) and Glucksberg (1993). Paradigmatic transparency refers to the "transparency of the semantic extension that leads from the original meaning of an expression to its transferred reading" (Geeraerts, p. 61). In this type of relation, the degree of derivability of the idiomatic meaning from the literal one depends on the availability of a motivating image. Examples of idioms with high and low paradigmatic transparency are "to keep a straight face" and "to hang fire," respectively. Transparency in the present study is operationally defined as the degree of overlap between the literal and the figurative meaning of an idiom, which comes closer to paradigmatic than to syntagmatic transparency.

The lack of overlap between literal and figurative meaning—the (sometimes only apparent) incongruity between the two—is one of the defining features of idiomatic expressions in general. Conventional definitions of the term *idiom* invariably make mention of the fact that idioms are fixed phrases whose overall (figurative) correct interpretation is different from the literal interpre-

tation of the sum of their constituent parts (e.g., Fraser, 1970). Early accounts of idiom structure and analysis (e.g., Weinreich, 1969) took an extreme position on the issue of idiom analyzability by claiming that idioms are noncompositional (i.e., that the meaning of the individual words does not contribute to the meaning of the idiomatic whole). Nunberg, Sag, and Wasow (1994), on the other hand, argued that idioms are compositional. This view has received support from a number of psycholinguistic studies such as Gibbs (1985, 1993) and Gibbs and Nayak (1989). In the context of these studies, the concept of transparency and especially its paradigmatic dimension in Geeraerts's (1995) terms have been investigated: Depending on the degree of semantic overlap or similarity between its literal and its figurative meaning, an idiom could be classified as transparent, opaque, or situated somewhere in between. This characteristic of idioms might be responsible for different learning and retention patterns with regard to comprehension and production, which is the reason why it is included in the current study.

According to Gibbs, Nayak, and Cutting (1989), idioms can be divided into three groups depending on their analyzability: normally decomposable, abnormally decomposable, and nondecomposable. The first group is comprised of idioms whose figurative meaning is related in a transparent way to their literal meaning (e.g., "to beat somebody at their own game"); the two are close enough semantically that the figurative meaning can be worked out on the basis of the composite literal meaning. Idioms belonging to the abnormally decomposable group display a less straightforward relationship between their literal and figurative readings (e.g., "to get off the hook"). Metaphorical mappings between the two, however, imbue these kinds of phrases with meaning. The resulting idiomatic meaning is related in a less transparent way to the literal meaning than is the case with normally decomposable idioms. The literal interpretation of nondecomposable idioms bears even less resemblance to their intended figurative meaning (e.g., "to paint the town red"). Such idioms are hardly transparent, unless one is able to trace them back to their possibly obscure origin or to some context that helped establish a certain culturally specific usage but is beyond recall for the lay language user and for L2 learners.

Culturally specific knowledge or the lack thereof might contribute to individual language users' perception of how analyzable idioms are. Bortfeld (2003) suggested that native speakers are guided by their "pre-established biases stemming from the phrases they have analyzed (or failed to analyze) from their own language" (p. 219) when they consider the degree of analyzability of L2 idioms. Therefore, the degree of analyzability (and thus transparency) necessarily remains a subjective measure, biased by the linguistic and cultural background of informants.

Irujo (1986) found that advanced L2 learners use their L1 knowledge to comprehend and produce L2 idioms. Idioms that were transparent, simple in terms of vocabulary and structure, and frequently used were the ones that were comprehended and produced most correctly. In a subsequent study, Irujo (1993)

asked Spanish learners of English to translate texts containing idioms into everyday English. No clear support was found for Kellerman's claim (1983) that the less transparent an idiom is, the less likely it is to be transferred from the L1 to the L2.

Transparency might be expected to have a greater facilitating influence on performance when receptive knowledge is tested. The more the literal and the figurative meanings of an idiom are felt to be related, the likelier it might be that one can decipher the figurative meaning of the idiom, basing one's interpretation on the clues that the literal meaning might make available. In the same vein, the literal meaning of a transparent idiomatic expression might also constitute a memory aid. A plausible interaction would be that performance is better for transparent idioms than for opaque ones, especially under L2-L1 testing conditions (i.e., that comprehension is easier for transparent idioms than for opaque ones). To illustrate this, imagine that one is given an idiom in the L2 and is required to provide a L1 equivalent or paraphrase: When the figurative, idiomatic meaning of the expression is substantially different from its literal meaning, starting off by considering the potential literal meaning might not be particularly helpful. In this line of reasoning, we propose two hypotheses related to the transparency of L2 idioms:

(3) Transparency hypotheses
   a. Performance on more transparent idioms will be better than performance on less transparent ones.
   b. The effect of transparency on performance will be stronger when receptive knowledge (L2-L1) is tested.

LONG-TERM RETENTION

One of the most important issues from a practical point of view in vocabulary acquisition is L2 learners' ability to recall words over long periods of time. A substantial amount of evidence suggests that when initial learning takes place under more difficult conditions, retention might be boosted in the long run (e.g., Battig, 1979; Jacoby & Craik, 1979; Schmidt & Bjork, 1992). Griffin and Harley (1996) claimed that the "depth of processing argument" (Craik & Lockhart, 1972), despite being "unsatisfactory as an explanation for learning (. . .) does have heuristic value" (p. 447): When initial learning takes place under more difficult conditions, it is possible that in order to compensate for this, participants engage in more varied and elaborate processing, which, in turn, helps establish stronger and more durable word-pair links.

The argument that more difficult learning would lead to better retention and less loss over time is not supported by De Groot and Keijzer's (2000) empirical findings. In a study that compared forward and backward translation (i.e., direction of testing; they did not investigate the effects of direction of learning), they found that words that were easier to learn were less susceptible to

forgetting in the period between the immediate and the delayed test than words that were difficult to learn. The title of their article encapsulates this finding in the words "what is hard to learn is easy to forget" (p. 1).

The lack of uniformity within studies that have applied the PAL paradigm makes it difficult to find a common conclusion in the literature. Griffin and Harley (1996) reported a decrease in performance over time, irrespective of learning condition. However, the authors focused on the difference between forward and backward association and, therefore, did not analyze which direction of testing would be more informative for the purpose of detecting time effects of direction of learning. Schneider et al. (2002) concluded that less loss over time was observed as a result of productive learning. Mondria and Wiersma (2004) focused on delayed performance because their immediate performance data showed a ceiling effect, which makes comparisons between immediate and delayed tests problematic.

Still, looking at the raw data (whenever available) instead of the conclusions each author drew from their respective data, it can be observed that delayed and immediate performance followed a very similar pattern. Performance varied to some extent on the receptive test depending on direction of learning, but much more so on the productive test. Performance on the productive test profited substantially from learning productively. This indicates that despite the natural forgetting over time, there remains a major advantage of productive learning when productive knowledge is required. Because we expect to find the same pattern of results on immediate performance for both idiom and word learning, it is interesting to investigate whether this pattern can also be found for delayed performance with regard to idiom learning. As such, two hypotheses are proposed for long-term retention effects:

(4) Delayed performance hypotheses
   a. Delayed performance on a productive test (i.e., L1-L2) will be better when idioms have been learned productively (L1-L2) rather than receptively (L2-L1).
   b. On a delayed receptive test (i.e., L2-L1), however, the performance of L1-L2 learners will not be superior to the performance of L2-L1 learners.

ASSESSMENT OF IDIOM IMAGEABILITY AND TRANSPARENCY

The design of the experiment required that we obtain independent imageability and transparency ratings for the 20 English idioms used in the main experiment.[1] Because the sample used in the main experiment consisted of university students who were L2 learners of English, we had to use a similar sample to gather the ratings. Thus, to collect data on the imageability and transparency of the stimulus material, we asked participants other than the ones who participated in the main experiment to rate the same idiomatic expressions in these two respects.

Method

Participants. Eighty native Dutch-speaking students (51 female, 29 male) aged 18–29 ($M = 21.5$, $SD = 2.46$) at the University of Amsterdam participated in the rating session, which was conducted as part of a 1-h test session at the Department of Psychology. Participants received €7 (approximately US$8) for their participation. University students in the Netherlands have typically had 6–7 years of English in high school; all participants could therefore be expected to have at least an intermediate level of proficiency in English.

Stimulus Material. Twenty idiom pairs (i.e., an English idiom and a Dutch equivalent; see Appendix) were used as stimulus material. The idioms were selected to fulfill three criteria. The main selection principle was frequency of the constituent words of the English idioms. Generally speaking, high-frequency words are more likely to be familiar to participants than low-frequency ones. Consequently, using expressions composed of high-frequency words guarantees that the form and meaning of the expressions as a whole rather than the form and meaning of the individual constituent words will constitute the focus of learning. The *Collins Cobuild English Dictionary*, which provides word-frequency information, was used to inform the choice of words used in the task. In this dictionary, word frequency is rated on a scale of 0–5 diamonds, which reflect the classification of words into six frequency bands (words with no diamonds are the least frequent in English). Only words with three to five diamonds were deemed acceptable constituents of the expressions to be used as learning and testing material in the study. The second criterion was the availability of a Dutch equivalent such that no one-to-one correspondence (in terms of literal translation) between the English and the Dutch constituent words of each idiom pair existed. The third criterion for selection was that the expressions had to be unfamiliar to intermediate and upper-intermediate nonnative speakers of English, so that these expressions could serve as targets in the learning experiment. Let us illustrate these criteria with the idiom "to hang fire": both "hang" and "fire" are high-frequency words, no one-to-one Dutch equivalent idiom (such as *vuur hangen*) exists, and it can be expected that this English idiom is not familiar to intermediate and upper-intermediate learners of English, such as the present sample.

Procedure and Instructions. Participants were given an English example of an idiom assumed to be familiar ("to pull someone's leg"). Subsequently, the example was used to illustrate the difference between the literal and figurative meanings of idiomatic expressions (in general), and participants were made aware of the fact that these two meanings could differ from each other to varying degrees in different idioms. Participants were then instructed to rate a list of idioms by indicating on a 7-point scale (1 = completely disagree, 7 = fully agree) their agreement with three statements regarding each of the 20 English idioms. In each case, along with the English idiom, a Dutch equivalent and a paraphrase in Dutch were given (see Appendix). Judgments with

respect to transparency ("The figurative meaning of this idiom has a lot in common with its literal meaning"), imageability ("I could easily visualize this idiom"), and participants' prior knowledge ("I already knew precisely what this idiom means") were elicited. Participants were neither instructed to learn the expressions they had to rate nor were they tested in any way subsequent to the rating. Idioms were presented in random order. Participants saw one idiom at a time on a computer screen and could not return to review or change their previous ratings.

Results

Table 1 summarizes the results of the study in the form of descriptive statistics. On the basis of participants' ratings of imageability, the 20 idioms were classified into three almost equally sized groups (low, intermediate, and high). Similarly, the 20 idioms were grouped into three classes according to their respective ratings for transparency.

THE LEARNING EXPERIMENT

As outlined previously, the first set of hypotheses are concerned with the influence of direction of learning on receptive and productive testing. The second and third sets of hypotheses are related to possible moderating effects of two idiom characteristics: imageability and transparency. The last two hypotheses address the effect of direction of learning on delayed performance.

Method

Participants. One hundred twenty-nine Dutch-speaking students[2] (96 female, 32 male) at the University of Amsterdam, aged 18–28 (M = 21.0, SD = 3.50), participated in the experiment, which was conducted as a part of a 1.5-h test session at the Department of Psychology. The students received €10 (approximately US$11) for their participation. University students in the Netherlands have typically had 6–7 years of English in high school; all participants could therefore be expected to have at least an intermediate level of proficiency in English.

Design. The experiment was conducted in a 2 × 2 × 2 design with two between-subjects factors—direction of learning (L1-L2 vs. L2-L1) and direction of testing (L1-L2 vs. L2-L1)—and one within-subject factor—time of testing. A list of 20 pairs of English idiomatic expressions and their Dutch equivalents was presented in the same order to all participants. Based on the ratings provided in the rating session, the English idioms differed in their degree of transparency (low vs. intermediate vs. high) and in their degree of imageability (low vs. intermediate vs. high). Idiom imageability and transpar-

Table 1. Rating measures of the 20 stimulus idioms

	Imageability			Transparency			Prior knowledge	
Idiom	*M*	*SD*	Classif.	*M*	*SD*	Classif.	*M*	*SD*
1. To get off the hook	4.94	1.81	Int.	4.55	1.80	High	4.95	2.13
2. To fly off the handle	4.11	1.84	Low	3.73	1.89	Int.	3.00	1.97
3. To lay something at somebody's door	5.21	1.66	High	4.43	1.83	High	4.00	2.04
4. To paint the town red	4.60	1.85	Int.	3.15	1.57	Low	3.24	2.09
5. To get cold feet (about something)	5.05	1.81	High	3.85	1.96	Int.	4.86	2.15
6. To hang fire	3.10	1.75	Low	2.52	1.47	Low	1.92	1.28
7. To stick to your guns	5.25	1.50	High	4.16	1.84	Int.	4.63	2.07
8. To have had your fill of something	4.98	1.86	Int.	4.87	1.54	High	4.61	2.20
9. To sit on the fence	4.59	1.81	Int.	3.62	1.68	Int.	2.88	1.88
10. To be in for it	3.43	1.77	Low	2.78	1.40	Low	2.90	1.81
11. To play the field	3.97	1.71	Low	3.43	1.67	Low	2.89	2.02
12. To show your hand	4.68	1.70	Int.	3.95	1.73	Int.	2.90	1.74
13. To wear your heart on your sleeve	5.35	1.65	High	4.15	1.91	Int.	4.73	1.95
14. To carry the day	3.76	1.85	Low	3.11	1.60	Low	2.90	1.80
15. To (manage to) keep a straight face	6.00	1.36	High	6.17	1.06	High	5.90	1.50
16. To have gone off the deep end	4.00	1.67	Low	3.40	1.54	Low	3.35	1.94
17. To put on airs	4.33	1.98	Int.	4.14	1.84	Int.	4.06	2.19
18. To beat someone at their own game	5.52	1.59	High	5.44	1.52	High	5.41	1.81
19. To be down in the dumps	5.26	1.43	High	4.89	1.61	High	4.65	1.92
20. To shoot/fire from the hip	4.33	1.88	Int.	3.27	1.61	Low	3.06	1.93

Note. Imageability, transparency, and prior knowledge ratings are assessed on a 7-point scale (1 = not at all; 7 = very much). Classif. = classification for the main experiment; Int = intermediate.

ency constituted two within-subject factors. To make sure that potentially present prior knowledge would not affect the experimental results, participants were randomly assigned to the four conditions that resulted from the between-subjects factors. Participants received all 20 stimulus idioms in a mixed order and were tested immediately after the learning session and then 3 weeks later. In the L1-L2 learning condition, the Dutch equivalent and a paraphrase in Dutch were presented first, followed by the English idiom. In the L2-L1 learning condition, the English idiom appeared on the screen first, followed by the Dutch equivalent and a paraphrase in Dutch.

Procedure and Instructions. It took participants approximately 25 min to complete the experiment, which was conducted in one uninterrupted session together with four unrelated psychological experiments. Upon arrival of the participants in the computer lab, the experimenter explained that all instructions would be given via the computer and then withdrew but remained nearby in order to be able to give assistance if necessary. Prior to the learning session, participants filled out a questionnaire eliciting proficiency-related information such as average secondary school grade in English and self-assessment of their current knowledge of English.

After completing the questionnaire, participants proceeded with the learning session. The instructions were given in Dutch. To start, participants were given an English example of an idiom assumed to be familiar to them: "to pull someone's leg." Subsequently, the example was used to illustrate the difference between the literal and the figurative sense of idiomatic expressions (in general), and participants were made aware of the fact that these two senses could differ from each other to varying degrees in different idioms. It was emphasized how important it is for foreign language learners to have some knowledge of idiomatic expressions for both comprehension and production purposes. The difficulty that foreign language learners often experience with idioms was addressed briefly, and it was highlighted that this kind of knowledge could contribute to nativelike fluency. The ensuing task was then presented as an opportunity to learn some more of these useful expressions.

In the instruction immediately preceding the learning task, participants were informed that the idiom pairs that they were supposed to learn would be presented one at a time in two successive presentations. Each pair remained on the screen for 30 s during the first round and 10 s during the second presentation round. A clock that remained on the screen above the text section throughout both sessions helped participants keep track of time. No mention was made of the variations in direction of learning and testing, and participants were not informed of the number of idioms they would be given to learn. Participants were instructed to try to learn the idioms as well as they possibly could. Several learning strategies were briefly suggested, such as paying attention to both the individual words used and the structure of the expressions as a whole, trying to think of situations in which the expressions could be used, and trying to form a mental image of the

expressions. Participants were told that they could use whichever strategy they preferred.

Before the second round of presentation, participants were reminded that the expressions would be presented at a faster rate than the first time and that it might help for them to recall what image (if any) they had associated with each expression and try to visualize it once again. During the second round of learning, the expressions were given in the same order and the direction of learning was the same as in the first round. Depending on the direction of the learning manipulation, the idiom pairs were presented either in the order of Dutch-English (the L1-L2 learning condition) or in the order of English-Dutch (the L2-L1 learning condition).

The test session began immediately after the second presentation of the idiom pairs. Depending on the direction of testing manipulation, participants had to type in either the Dutch equivalent or the Dutch paraphrase (both given during the learning session) of the English idioms learned (the L2-L1 testing condition), or the English equivalent of the Dutch idioms (the L1-L2 testing condition). Following the test session, participants were asked to indicate on a 7-point scale (bounded by 1 = completely disagree, and 7 = fully agree, with the midpoint 4 = neither agree nor disagree) their agreement with four statements with respect to each of the English idioms. Perceived ease of learning ("I found it easy to commit this idiom to memory"), transparency ("The figurative meaning of this idiom had a lot in common with its literal meaning") and imageability of the idioms ("I could easily visualize this idiom"), and participants' prior knowledge ("I already knew precisely what the idiom means") were thus assessed. Finally, participants could sign up for feedback (by entering their e-mail address).

Delayed Test. Three weeks after the last session of the main experiment, a delayed test was conducted in an attempt to collect data on the retention of idiomatic expressions over time. This time the participants were approached by e-mail and were requested to e-mail back their answers to the test. They were given a short explanation of the purpose of the follow-up test along with the instructions. The format of the test was the same as in the main experiment and all participants were tested in the same direction as in the first test. Fifty-eight participants responded (45.0% of the original sample). The majority of them responded to the first e-mail ($n = 44$). A second e-mail was sent to those who had not responded within 8 days, and another 14 participants responded as a result. All participants received feedback via e-mail, including the list of expressions they were tested on and some suggestions for online resources on idioms.

Scoring. The English answers (i.e., in the L1-L2 testing condition) were scored on a 5-point scale in accordance with two basic rules: (a) One point was awarded for each correct stem of any of the content words that occurred in the idiom (the maximum number of points that could be awarded with respect to this criterion was three) and (b) one point was given for correct

use of prepositions (including the absence of prepositions in some examples), plural versus singular form of nouns, and other elements of the phrase (i.e., aspects that relate to the accuracy with which the overall phrase structure was produced). As a rule of thumb, when all content words and prepositions were used correctly but the use of singular versus plural forms or the use of articles deviated from the original, one point was subtracted from the maximum. The smaller the number of words in an idiom, the more such deviations weighed. Spelling mistakes were ignored in all cases.

To illustrate these scoring rules, we will now consider several of the actual responses subjects gave on the test when they were supposed to produce the idiom "to stick to your guns." No points were given for "stubborn" and "don't move your hips." One point was awarded to, for instance, "to stick with," "to hold the guns," "to stick by," or "to stick on your. . . ." Examples of responses that received two points were "to stay with your guns," "to stick to it," and "to plug to your gun," whereas "to stay to your guns," "to stick to your own gun," "to stick your guns," and "to hold to your guns" were awarded three points. Finally, four points were given to "sticking to your guns," "to stick to ya guns," and, of course, "to stick to your guns."

A 5-point scale was also used for the scoring of the Dutch answers. No answer or a wrong answer were scored as zero, a correct answer (be it the Dutch idiom suggested as an equivalent of the expression to be learned or the complete paraphrase) was awarded four points, and there were three levels of approximation of meaning in between, depending on the extent to which different aspects of the meaning were correctly and fully reflected in the response.

Two raters independently assigned performance scores to the responses of 129 participants in the immediate test (2580 scores) and of 58 participants in the delayed test (1160 scores). Of the 3740 scores that each rater assigned in total, the two raters assigned identical scores in 3115 cases (83.3%). To judge the interrater reliability, four intraclass correlation coefficients (ICCs) were calculated for the ratings of each idiom, separately for the two directions and the two moments of testing. In one case (immediate receptive performance on idiom 4), both raters assigned the same score to all participants ($n = 57$). In the remaining cases, 65 ICCs were greater than .80 and nine ICCs were above .60. Only five ICCs were below .60 and therefore pointed to low interrater reliability, but three of them referred to delayed performances in the receptive test. Because of this high agreement between the raters, the scores of the first rater were used for subsequent analyses.

Dependent Measures. Total immediate performance was calculated by adding up the performance scores for all 20 idioms (which resulted in a total maximum score of 80). Immediate performance on the low, intermediate, and high imageable idioms was calculated by averaging the respective performance scores of the idioms that were ranked as low (six in total), intermediate (seven in total), and high (seven in total) on imageability in the norming study (see

Table 1). Immediate performance on the low, intermediate, and high transparent idioms was calculated by averaging the respective performance scores of the idioms that were ranked as low (seven in total), intermediate (seven in total), and high (six in total) on transparency in the norming session. In the case of delayed performance scores, collected via e-mail, sum scores of total delayed performance, delayed performance on low, intermediate, and high imageable idioms, and delayed performance on low, intermediate, and high transparent idioms were calculated in the same way.

Results

Descriptive Statistics. Table 2 presents the immediate and delayed performance scores on the 20 idioms and their ratings of ease of learning, imageability, transparency, and prior knowledge. In the rightmost column of Table 2, it is apparent that, overall, participants indicated that their prior knowledge of the 20 stimulus idioms was rather low: The reported means were above the scale midpoint 4 for only two idioms. The means reported in Table 2 therefore indicate that participants rejected the statement "I already knew precisely what the idiom means" in 18 out of 20 cases.

Effect of Direction of Learning and Direction of Testing on Immediate Performance. The first set of hypotheses predicted that direction of learning would affect immediate performance, especially for the productive test. Hypothesis 1a predicted that performance on the productive test (L1-L2) would be better when idioms were learned productively (L1-L2) rather than receptively (L2-L1). According to hypothesis 1b, however, productive learning was not expected to improve performance on the receptive test (L2-L1).

The results of the present study are summarized in Table 3. Comparing these results with the results of earlier studies on paired-associate word learning gives initial support to our expectations. As in all prior studies, a large difference in performance between the two directions of learning was observed in the productive test ($M = 52.77$ vs. $M = 33.27$), whereas the difference between the two directions of learning observed in the receptive test was smaller in size ($M = 60.22$ vs. $M = 68.30$). However, it should be noted that this difference cannot be analyzed in terms of statistical significance due to the fact that the comparison is between two different scales (one for the L1-L2 responses and one for the L2-L1 responses). These two scales, although bounded by the same numbers, reflect different kinds of performance. Due to the similar range and division of the scales, however, a careful comparison should be possible as long as its result is considered indicative of differences rather than as proof of differences.

To test hypothesis 1a, we compared immediate performance on the productive test of participants who learned receptively with the performance of participants who learned productively. The *t* test revealed that participants who had learned productively did indeed perform significantly better than par-

Table 2. Performance and rating measures of the 20 stimulus idioms

	Immediate performance				Delayed performance				Ease of learning		Imageability		Transparency		Prior knowledge	
	L1-L2 (n = 72)		L2-L1 (n = 57)		L1-L2 (n = 25)		L2-L1 (n = 33)		(n = 128)		(n = 128)		(n = 128)		(n = 128)	
Idiom	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	2.35	1.71	2.72	1.78	0.84	1.46	1.79	1.69	4.34	2.0	4.13	1.9	4.38	1.6	3.44	2.1
2	1.14	1.61	2.21	1.91	0.80	1.47	2.18	1.81	3.66	1.9	3.80	1.9	3.85	1.7	2.51	1.5
3	2.69	1.43	3.63	1.05	1.64	1.66	3.09	1.59	5.27	1.6	5.28	1.5	5.12	1.4	3.71	2.0
4	3.58	1.00	4.00	0.00	3.52	1.33	3.82	0.77	5.86	1.3	5.72	1.4	4.17	1.8	3.48	2.1
5	3.15	1.54	2.88	1.50	0.68	1.49	1.27	1.15	5.04	1.8	4.59	1.7	4.01	1.7	3.78	2.1
6	1.92	1.73	2.79	1.84	0.60	1.35	0.82	1.55	3.54	2.0	3.19	1.8	2.74	1.4	2.36	1.6
7	1.63	1.72	3.25	1.18	1.56	1.83	2.15	1.68	4.78	2.0	4.59	1.9	4.20	1.7	3.46	2.1
8	1.19	1.51	3.93	0.53	1.32	1.68	3.88	0.70	4.23	2.0	4.05	1.9	4.48	1.7	3.38	2.1
9	2.71	1.67	3.47	1.28	1.80	1.94	1.73	1.89	4.81	1.9	5.12	1.7	4.32	1.7	3.13	1.9
10	2.08	1.84	2.70	1.73	0.60	1.41	0.55	1.25	3.73	2.0	3.02	1.7	3.02	1.6	2.84	1.9
11	2.51	1.73	3.77	0.91	1.32	1.84	1.76	1.97	4.97	1.9	4.99	1.7	4.36	1.7	3.33	2.0
12	2.89	1.51	3.09	1.46	2.16	1.46	0.88	1.29	5.05	1.7	5.14	1.7	4.84	1.6	3.23	2.1
13	2.15	1.59	3.23	1.28	0.60	1.35	2.00	1.37	5.02	1.8	5.26	1.7	4.82	1.7	3.64	2.2
14	2.10	1.89	3.12	1.65	0.68	1.41	2.12	1.93	4.39	2.0	4.00	1.9	3.73	1.8	2.91	1.9
15	2.65	1.61	3.67	0.87	3.00	1.41	2.88	1.41	5.54	1.6	5.56	1.5	5.70	1.4	4.90	2.1
16	1.31	1.52	2.79	1.81	0.52	1.33	1.79	1.95	3.93	2.0	4.16	1.8	3.83	1.6	3.03	1.9
17	1.65	1.39	3.46	1.27	1.12	1.39	2.52	1.66	4.98	1.8	4.99	1.6	4.41	1.7	3.75	2.0
18	1.93	1.77	3.79	0.65	1.16	1.75	3.39	1.06	5.40	1.6	5.25	1.5	5.28	1.4	4.36	2.2
19	1.36	1.56	3.23	1.59	1.24	1.51	2.97	1.61	4.59	2.0	4.79	1.7	4.49	1.7	3.38	2.0
20	1.75	1.84	2.75	1.76	0.72	1.49	1.12	1.52	4.22	1.9	4.27	1.9	3.54	1.7	2.73	1.7

Table 3. Total immediate and total delayed performance by experimental condition

Test by group and measure	*n*	*M*	*SD*	Range
English-Dutch (L2-L1)				
L2-L1 learning group				
Immediate performance	30	68.30	10.66	37–80
Delayed performance	18	45.28	16.04	17–69
L1-L2 learning group				
Immediate performance	27	60.22	14.33	30–80
Delayed performance	15	39.60	14.11	19–62
Dutch-English (L1-L2)				
L2-L1 learning group				
Immediate performance	37	33.27	14.83	6–62
Delayed performance	11	19.36	14.02	4–49
L1-L2 learning group				
Immediate performance	35	52.77	17.89	21–80
Delayed performance	14	31.00	18.36	2–61

Note. The maximum possible score is 80. Different scales were used for scoring the Dutch and English responses. Thus, although for both tests the maximum score is 80, the figures in the English-Dutch group cannot be compared with those in the Dutch-English group.

ticipants who had learned receptively, $t(70) = 5.05$, $p < .001$. To test hypothesis 1b, the same comparison was made for immediate performance on the receptive test. Confirming the prediction that L1-L2 learners would not outperform L2-L1 learners, the *t* test revealed that participants who had learned receptively performed even better than participants who had learned productively, $t(55) = 2.43$, $p < .05$.

The reasoning that led us to propose this first set of hypotheses, however, also suggests that direction of learning would be especially relevant in the more demanding productive test. To address this, we compared the magnitude of the effects found for both hypotheses. We calculated the two respective effect sizes by dividing the difference between the productive learners and the receptive learners (as displayed in Table 3) by the overall standard deviation on each test. The effect size of the finding related to hypothesis 1a is $(52.77 - 33.27)/19.00 = +1.03$. That means that on the productive test, the productive learners outperformed the receptive learners by more than one standard deviation, which can be considered a very large effect. The effect size of the finding related to hypothesis 1b is $(60.22 - 68.30)/13.06 = -0.62$. This effect size is negative, which shows that productive learners performed worse than receptive learners on the receptive test. More importantly, this second effect size is considerably smaller than the first: Receptive learners outperformed productive learners by less than two thirds of a standard deviation. All in all, the results support the first set of hypotheses: Direction of learning affected immediate performance, and its effect was particularly large on the productive test.

Moderating Effects of Imageability and Transparency. The second and third sets of hypotheses addressed potential moderating effects of two idiom characteristics: imageability and transparency. For reasons of efficiency, we first report on the findings with respect to hypothesis 2c and then on those concerning hypotheses 2a and 2b. To test hypothesis 2c, predicting that high positive correlations would be observed between performance and the rating of one's ability to visualize the idioms to be learned, we calculated three sets of correlation scores.

The first set of correlations, displayed in the first row of Table 4, is based on the mean immediate performance scores (collapsed over both tests) and ratings over all 20 idioms. Performance is significantly and positively correlated with both ratings, which shows that participants who performed better rated it less difficult to visualize the idioms and considered the idioms to be more transparent. We compared the two correlation coefficients (i.e., the correlation between immediate performance and imageability and the correlation between immediate performance and transparency) using the formula proposed by Olkin and Siotani (1964, in Bortz, 1989) to find out whether the first correlation is significantly higher than the second one. As shown in the third and fourth columns of Table 4, this was the case;[3] that is, although performance is significantly correlated with both ratings, the relation between performance and imageability is even stronger than the relation between performance and transparency. The remaining rows of Table 4 (rows 2–7) break down this general finding for each separate class of idioms. Recall that the subcategorization into three classes of imageability and three classes of transparency was based on the ratings of imageability and transparency provided in the preliminary assessment study, independent of the ratings by students

Table 4. Correlation of mean immediate performance scores and mean ratings of imageability and transparency summarized per class of idioms

			Comparison[a]	
Idioms by ranking	*r*(IP + imageability)	*r*(IP + transparency)	*z*	*p*
All 20 idioms	.55****	.41****	2.63	***
Imageability				
Low	.51****	.29****	3.36	****
Intermediate	.51****	.29****	2.05	**
High	.51****	.41****	1.75	*
Transparency				
Low	.47****	.26***	2.92	***
Intermediate	.58****	.38****	3.39	****
High	.47****	.49****	−0.44	*ns*

Note. IP = Immediate performance. $N = 128$.
[a]Comparison of the two correlation coefficients in columns 2 and 3.
*$p < .10$. ** $p < .05$. *** $p < .01$. **** $p < .001$. All p's are two-tailed.

in the learning experiment. For each subgroup of idioms, performance is substantially and positively correlated with both ratings. This lends additional support to the finding that participants who performed better rated it easier to visualize the idioms and considered the idioms to be more transparent than participants who did not perform as well.

Hypotheses 2a and 2b predicted that imageability would facilitate immediate performance, especially when the idioms were learned in the L2-L1 direction. In order to analyze these predictions, we compared three dependent variables per participant; that is, their mean performance score on the six low imageable idioms, the seven intermediate imageable idioms, and the seven high imageable idioms. In terms of multivariate analysis of variance (MANOVA), this adds the three-level within-subject factor imageability to the experimental design.

As the second set of hypotheses predicted within-subject effects, a $2 \times 2 \times 3$ MANOVA with the between-subjects factors of direction of learning and direction of testing and the within-subject factor of imageability was performed. Note that this analysis does not compare performances on different tests (receptive vs. productive) but different performances within each participant's dataset. Thus, finding an interaction that involves the between-subjects factor direction of testing would not indicate that performances on the different tests differ, but it would indicate that direction of testing differentially influenced participants' performance on low, intermediate, and high imageable idioms. Still, the comparison is between performances on identical tests because performances are compared within each participant. The thorny issue of whether recognition is an easier test than production is not central to this study, as comparing different tests yields complications in terms of the scoring of the performance.

To test hypothesis 2a, the mean performance scores on low, intermediate, and high imageable idioms were entered into a $2 \times 2 \times 3$ MANOVA with direction of learning (L2-L1 vs. L1-L2) and direction of testing (L2-L1 vs. L1-L2) as between-subjects factors and imageability (low vs. intermediate vs. high) as a within-subject factor. A main effect was obtained for imageability, $F(2, 250) = 31.42$, $p < .001$. Paired-samples t tests revealed that participants performed worse on low imageable idioms ($M = 2.31$, $SD = 1.22$) than on intermediate ($M = 2.76$, $SD = 0.98$), $t(128) = -6.70$, $p < .001$, or high imageable idioms ($M = 2.74$, $SD = 1.03$), $t(128) = -6.11$, $p < .001$, whereas performance on the latter two did not differ, $t(128) = 0.53$, $p = .60$. MANOVA further revealed the predicted interaction between imageability and direction of learning, $F(2, 250) = 3.61$, $p < .05$. As predicted by hypothesis 2b, imageability had a weaker moderating effect on the performance of participants who learned in the L1-L2 direction ($M = 2.55$, $SD = 1.10$; $M = 2.84$, $SD = 0.88$; and $M = 2.97$, $SD = 0.80$ for low, intermediate, and high imageable idioms, respectively) than on the performance of participants who learned in the L2-L1 direction ($M = 2.08$, $SD = 1.28$; $M = 2.69$, $SD = 1.07$; and $M = 2.52$, $SD = 1.18$, respectively). This finding is illustrated in Figure 1.

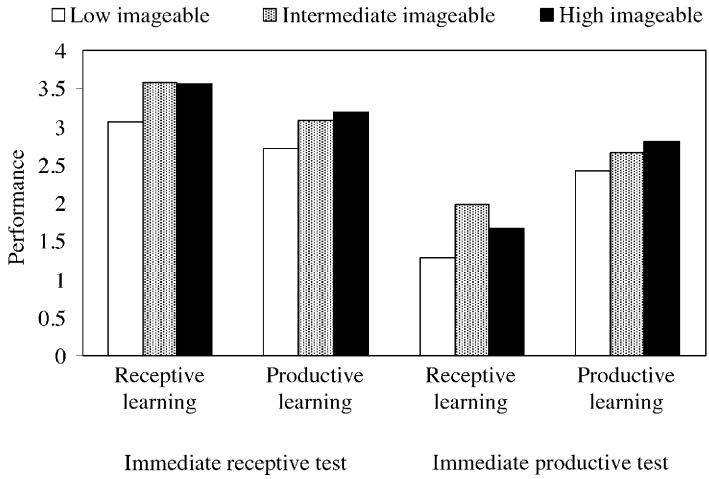

Figure 1. Immediate performance depending on condition and imageability.

There was no significant interaction with direction of testing, both with $F(2, 250) < 1$. This finding shows that the patterns predicted by hypothesis 2a and 2b and found in this analysis were the same in the receptive and the productive tests. The fact that no interaction involving direction of testing was observed indicates that this advantage can be found on both tests (productive and receptive). For productive learners, there was gradual improvement in performance across the three groups of idioms that differed in imageability. For receptive learners, however, there was a steep (and significant) increase in performance between the low and the intermediate imageable idioms, whereas the difference between performance on the intermediate and the high imageable idioms was not significant (i.e., it seemed especially inefficient to learn low imageable idioms in the L2-L1 direction). These findings confirm and qualify the conclusions that we made about the role of imageability based on the correlations between imageability ratings and performance (as predicted in hypothesis 2c). Altogether, the second set of hypotheses received good empirical support.

The third set of hypotheses predicted that transparency would facilitate immediate performance, especially when receptive knowledge was tested. Analogous to the procedure described previously, we added the three-level within-subject factor transparency to the experimental design. In hypothesis 3a, we predicted that test performance on more transparent idioms would be better than that on less transparent ones, and in hypothesis 3b, we predicted that this effect would be particularly pronounced when receptive knowledge was tested (L2-L1). To test these hypotheses, the mean performance scores on low, intermediate, and high transparent idioms were entered into a MANOVA sim-

ilar to the one described for the imageability results. Results revealed that transparency had no significant main effect on performance, $F(2, 250) = 2.32$, $p = .10$. MANOVA did reveal, however, the interaction between transparency and direction of testing, $F(2, 250) = 11.81$, $p < .001$. Figure 2 shows that, as predicted by hypothesis 3b, high transparent idioms were comprehended particularly well (i.e., in the L2-L1 testing direction, with $M = 3.13$, $SD = 0.86$; $M = 3.08$, $SD = 0.75$; and $M = 3.49$, $SD = 0.60$, for low, intermediate, and high transparency idioms, respectively), whereas transparency had no facilitating effect on production ($M = 2.18$, $SD = 1.07$; $M = 2.19$, $SD = 1.04$; and $M = 2.03$, $SD = 1.05$, respectively).

Long-Term Retention. The fourth set of hypotheses predicted that direction of learning would affect delayed performance, especially for the productive test.[4] Hypothesis 4a predicted that performance on the delayed productive test (i.e., L1-L2) would be better when idioms were learned productively (L1-L2) rather than receptively (L2-L1). According to hypothesis 4b, however, productive learning was not expected to improve delayed performance on the receptive test (i.e., L2-L1). The difference in delayed performance between the two directions of learning that was observed for the productive test ($M = 31.00$ vs. $M = 19.36$) was greater than the difference between the two directions of learning observed for the receptive test ($M = 39.60$ vs. $M = 45.28$, see also Table 3). To test hypothesis 4a, we compared the delayed performance on the productive test of participants who learned receptively with the delayed performance of those who learned productively. Confirming hypothesis 4a, the *t* test revealed that participants who learned productively performed better

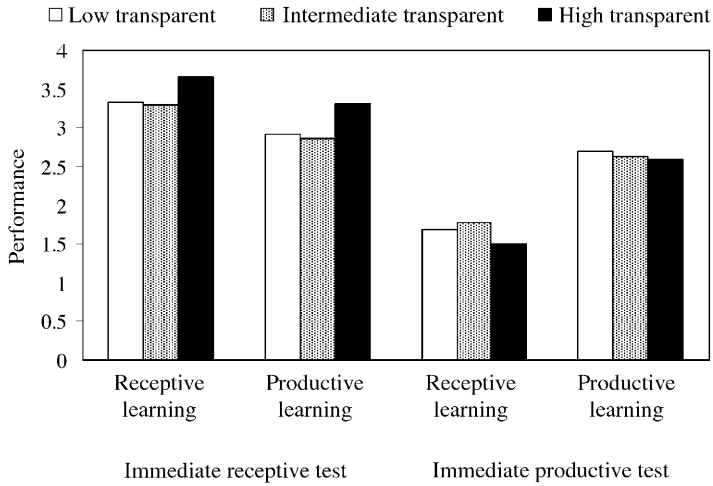

Figure 2. Immediate performance depending on condition and transparency.

than participants who learned receptively, $t(23) = 1.74$, $p < .05$ (one-tailed). To test hypothesis 4b, the same comparison was made for delayed performance on the receptive test. Confirming the prediction that L1-L2 learners would not outperform L2-L1 learners, the *t* test revealed that direction of learning did not influence delayed performance, $t(31) = 1.07$, $p = .29$.

Comparing the effect sizes for the delayed tests, again, supports the reasoning that led us to propose hypotheses 4a and 4b. The effect size of the finding related to hypothesis 4a is $(31.00 - 19.36)/17.30 = +0.67$, which is to say that even 3 weeks after learning, productive learners still outperformed receptive learners on the productive test by more than two thirds of a standard deviation, which is a considerable effect. The effect size of the finding related to hypothesis 4b is $(39.60 - 45.28)/15.23 = -0.37$, which is small, as no differences between the groups were found. All in all, the results support the fourth set of hypotheses: Direction of learning affected delayed performance, but only on the productive test.

To examine the issue of decline in test scores over time, we performed an explorative analysis and submitted immediate and delayed performance scores into a $2 \times 2 \times 2$ repeated-measurements MANOVA with direction of learning (L2-L1 vs. L1-L2) and direction of testing (L2-L1 vs. L1-L2) as between-subjects factors and time of measurement (immediate vs. delayed test) as a within-subject repeated-measures factor. This analysis is based on 45% of the sample, because delayed performance scores could be obtained from only 58 participants.

A main effect of time, $F(1, 54) = 191.58$, $p < .001$, indicated that, overall, immediate performance scores were higher than delayed performance scores ($M = 59.07$, $SD = 18.53$ vs. $M = 35.45$, $SD = 18.08$). This deterioration was qualified over time by two interactions. An interaction of the repeated-measures factor with direction of learning, $F(1, 54) = 4.80$, $p < .05$, showed that deterioration followed a different pattern depending on the direction of learning. Performance scores of participants in the L1-L2 learning direction decreased more over time (i.e., from $M = 62.21$, $SD = 13.85$ to $M = 35.45$, $SD = 16.59$) than did those of participants in the L2-L1 learning direction (i.e., from $M = 55.93$, $SD = 22.06$ to $M = 35.45$, $SD = 19.75$). A three-way interaction of the repeated-measures factor with direction of learning and direction of testing, $F(1, 54) = 5.98$, $p < .02$, showed that there were differences in the pattern of deterioration across the four experimental conditions. As Figure 3 illustrates, different patterns of results emerge for L1-L2 and for L2-L1 testing. Although recognition scores dropped by virtually the same amount in both direction-of-learning subgroups (the first and second pairs of bars from the left), direction of learning affected the decline in production performance differently: Production scores of participants in the L2-L1 learning condition (fourth pair of bars from the left) decreased less than scores of participants in the L1-L2 learning condition (fifth pair of bars from the left), although, as found in the analysis related to hypothesis 4b, the latter still scored better on the delayed test, $t(23) = 1.74$, $p < .05$ (one-tailed).

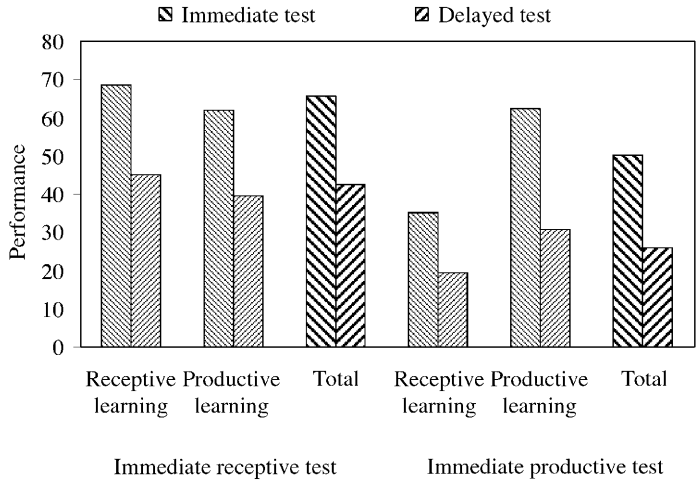

Figure 3. Immediate and delayed performance depending on condition.

GENERAL DISCUSSION

In this study, L2 idiom learning was examined in a PAL task to investigate the effects of direction of learning, direction of testing, idiom imageability, idiom transparency, and long-term retention. In our discussion, we will recap the main results of the study, situate the implications of our findings in a broader context, and make some suggestions for further research.

L2 Idiom Learning and Direction of Learning

In an attempt to extrapolate from the findings regarding paired-associate studies on the learning of individual, isolated, L2 words (Griffin & Harley, 1996; Mondria & Wiersma, 2004; Schneider et al., 2002; Stoddard, 1929), we used L2 idioms as stimulus material in order to explore whether similar principles hold when participants learn combinations of familiar words whose meaning as a whole is not familiar. We based our predictions on the idea that learning vocabulary under conditions that make the learning procedure itself more difficult (i.e., L1-L2) would lead to more processing and more remembering (Schneider et al., 2002) and on the principle of transfer-appropriate processing (Bransford, Franks, Morris, & Stein, 1979), which states that a processing compatibility between learning and testing tasks will positively affect testing task performance, whereas incompatibility will negatively affect testing task performance.

Confirming this line of reasoning, we found that (a) performance on the productive test was influenced differentially by direction of learning, in such

a way that performance was better when, prior to the test, idioms had been learned productively, whereas (b) productive learners did not outperform receptive learners on the receptive test. The first finding suggests that the productive direction of learning mitigates the difficulty of the productive test, whereas the receptive direction of learning does not prepare participants well for the doubly demanding backward testing. The pattern of results was different on the receptive test: A comparison between the performance scores of the two groups of learners showed that participants who had learned receptively performed significantly better than participants who had learned productively.

These findings are interesting for at least two reasons. On the one hand, they contribute to the literature on PAL, as they suggest that the underlying principles that might be responsible for the differential pattern of results (depending on direction of learning) in the case of word learning appear to influence idiom learning in the same way, despite the fact that learning involves establishing different types of association in these two cases. On the other hand, a notable aspect of the findings is their potential practical implication for productive performance, the major deficiency in L2 learners' proficiency with respect to idiomatic expressions being more on the productive than on the receptive front. As the main goal of most L2 learners is to apply their L2 knowledge in real-life situations, it is imperative that they tackle the greater hurdle of production as efficiently as possible.

Mondria and Wiersma (2004) suggested that the productive direction of learning triggers processing that is different in some way (quantitatively, qualitatively, or both) from the processing triggered by receptive learning. Quantitative differences would imply different depth of processing, whereas qualitative differences might have to do with the establishment of different kinds of associations between form- and meaning-related aspects of the L1 and the L2 expression. However, despite the impact that depth of processing has had on theoretical thinking (Lockhart, 2002), it is still a concept that is notoriously difficult to measure, so steering a depth-of-processing course of investigation should only be done after careful consideration of this caveat. The present findings tie in with the idea of depth of processing as explored thus far in the literature. We tentatively suggest that this be explored in future research, as our data were not collected with a specific emphasis on this issue.

One possible way to approach this and other processing-related questions in further research might be to collect data on the online processing of L2 idioms by asking participants to report the strategies they applied during comprehension and learning in a think-aloud session. Cooper (1999), who employed this approach for a comprehension study that presented L2 idioms to learners in a rich context, concluded that "use of context was the major strategy employed by the participants to arrive at the meaning of the expressions" (p. 258). He also suggested that it might be fruitful to present L2 idioms "in a nonsupportive context" to see what other strategies might play a role when L2 learners try to comprehend and learn L2 idioms (p. 258). A more recent

study by Liontas (2002) in this vein attempted to access online processing information by, among other things, letting L2 learners report on the strategies they applied while trying to explain the idiomatic meaning of L2 idioms without relying on any supporting context. This yields a wealth of meaning-making strategies, which could potentially enrich our understanding of the mechanisms involved in the L2 processing of idioms.

Imageability and Dual Coding

To further explore some language-specific cognitive mechanisms that potentially qualify the differential direction of findings in PAL, we investigated two characteristics of idioms: imageability and transparency. We obtained evidence that imageability and performance were positively correlated and, more interestingly, the receptive and the productive directions of learning produced different patterns of results along the dimension of imageability. The finding that direction of learning differentially affected the efficiency with which idioms that differed with respect to their degree of imageability were learned is the most interesting qualification of the results. Performance scores of participants who learned receptively were affected in a different way than performance scores of participants who learned productively. This suggests that the processes taking place during receptive learning are possibly more sensitive to differences in imageability than are the processes that take place during productive learning.

A potential complication of the issue of imageability and its implications is inherent in the question of whether the literal or the figurative meaning of idioms is normally visualized. Cacciari and Glucksberg (1995) investigated the question of whether the images evoked by L1 idioms are based on the literal or figurative reading and found that the literal rather than the more abstract figurative meaning was usually reflected in the respective images. Cacciari and Glucksberg claimed that "for unfamiliar idioms, especially those whose meanings are not known at all, only literal images should be possible" (p. 47). If this also holds true with regard to learners' processing of unfamiliar L2 idioms, then low imageability should, in effect, go hand in hand with greater difficulty of forming an image of the literal meaning.

There could be alternative explanations for the finding that direction of learning differentially influences performance on idioms that differ in imageability. It could be the case that during receptive learning (i.e., L2-L1), more emphasis is laid on the literal wording because the first stage of learning involves an encounter with an unfamiliar expression. This might trigger a standard so-called autopilot strategy of handling the situation by analyzing the syntax of the structure and referring to the meaning representations of the individual constituents, the result of which can be worked out as a literal semantic interpretation. Only then can the figurative meaning (provided a few moments later) be attached to the initial literal interpretation. During produc-

tive learning, on the other hand, it could be the case that more emphasis is placed on the conceptual representation of the L2 idiom than on its form. Also, the learner might analyze the degree of correspondence between the conceptual representations of the L1 and the L2 idiom.

At this stage it might be premature to draw parallels with established models of the bilingual lexicon when considering the nature of the L1-L2 form-meaning associations and the role of imageability in the case of L2 idioms. At present, little is known about the structure of the bilingual lexicon with regard to figurative language and multiword units (Kroll & Tokowicz, 2005). Our experiment was not designed to investigate what the L2 idiom entries in the bilingual lexicon look like (i.e., whether there are special idiomatic entries for known L2 idioms or whether the meaning of L2 idioms has to be computed after analyzing the components of the literal meaning and any relevant background information). More research is needed to fill this gap, as models of L1 idiom comprehension and processing cannot be readily applied to L2 idioms.

L2 Idiom Learning and Direction of Testing

Comprehension is consistently estimated to be easier than production. A number of researchers have been unanimous in their acceptance of this recurrent finding (e.g., Horowitz & Gordon, 1972; Kroll, 1993; Mägiste, 1979). According to Kroll, one possible explanation is that the translation direction used in a comprehension task (L2-L1) is the same as the direction in which learners are usually familiarized with unknown L2 words—namely the unknown L2 word first, followed by the L1 translation. As Griffin and Harley (1996) put it, in comprehension one is "working toward the well known; if the provision of the L2 cue (. . .) results in even a minimal activation of memory, then participants are in a position to choose between candidate responses, which will be real L1 (. . .) words," but "in the case of production, participants are working from the known to the less well known or even the unknown" (p. 446).

As far as idioms are concerned, it should be noted that certainly not all L2 idioms present learners with great difficulties, at least as far as comprehension is concerned. Sometimes an unknown L2 idiom makes sense to learners despite the fact that they have never heard or seen that idiom previously. In a study on crosslinguistic comprehension of idioms, Bortfeld (2003) demonstrated that, remarkably, even people who did not have any knowledge whatsoever of a certain language could correctly comprehend idioms from that language when asked to judge the respective literal English (L1) translations and categorize them conceptually. Bortfeld suggested that all people across languages and cultures share similar embodied experiences that form the basis for schematic representations, which, in turn, "motivate and structure how specific phrases—such as idioms—evolve in any single language" (p. 227) and guide the comprehension of these phrases across languages. This line of reasoning could be advanced as an additional explanation for the ease of com-

prehension of idioms as compared to production; that is, from a certain level on, once the constituent words of L2 idioms are familiar, learners will be in a position to provide more educated guesses than random guessing when comprehension is tested, which might render even an unfamiliar L2 idiom comprehensible.

In other words, selecting only widely unknown idioms might not be enough to block all possible sources of knowledge and intuitions. Even without prior knowledge, performance on a receptive test can result from processes other than learning alone (e.g., guessing). This is an alternative explanation for the seemingly better performance on the receptive test (relative to the productive test). The fact that participants can perform well as a result of guessing rather than as a result of learning might also partly account for the finding that the results were less pronounced in the receptive testing condition. Guessing did not affect the main findings in this study, however, because the results revealed that the experimental manipulations had differential effects within the productive test (and, to a lesser degree, also within the receptive test).

Transparency as a Facilitator of Recognition

Transparency—the degree of semantic overlap or similarity between the literal and the figurative meaning of an idiom—is another characteristic that was expected to qualify the findings on L2 idiom learning. Contrary to our expectations, performance scores did not increase in proportion to the degree of transparency. This can be interpreted as disproving the idea that the greater the overlap between literal and figurative meaning, the better the literal meaning could serve as a memory aid. Similarly, it could be surmised that a greater overlap between literal and figurative meaning would result in L2 idioms being understood better (or being more readily considered as motivated), which might, in turn, lead to better retention, but our empirical results do not substantiate such a prediction. However, as predicted, transparency did have differential effects on performance depending on the direction of testing. Hardly any variation in performance scores due to differences in transparency was observed on the productive test. However, when comprehension was tested, performance was actually boosted by higher transparency, which substantiates the hypothesis concerning differential effects of transparency on performance in the different test conditions. High transparent idioms were comprehended better than low or intermediate ones, whereas higher transparency did not have a similar facilitating effect on production.

A Comparison of the Predictive Potential of Imageability and Transparency

We will now discuss some findings based on correlations between performance in the main experiment and the ratings of imageability and transpar-

ency that participants provided after the test. Imageability was found to be strongly and positively correlated with immediate performance. Positive correlations were also found between immediate performance and transparency, but they were not as strong as those between performance and imageability. This difference could even be demonstrated to be statistically significant and in favor of imageability, which suggests that imageability might be a better predictor of performance than transparency. The correlational findings suggest, on the one hand, that participants who reported post hoc that they found it easy to conjure up an image during learning performed well during testing. On the other hand, performance scores improved with increasing imageability of the idioms (as rated in the norming study). These observations might be interpreted as initial confirmation of the proposed relevance of dual coding (Paivio, 1969) in idiom learning, as high imageable idioms (i.e., those most likely to evoke a mental image and be encoded along that dimension as well) were indeed learned better in all conditions. Importantly, the mental image evoked can only be expected to increase performance when it has been generated with relation to the L2 idiom because it is the L2 idiom that can profit from this memory aid.

The finding that higher transparency facilitated comprehension is neither surprising nor very impressive, as it only suggests that learners can more easily grasp the figurative meaning of an idiom by using the clues offered by the literal meaning when the two are rather similar anyway. More importantly, however, the finding that productive performance was not affected proportionately by increasing transparency suggests that transparency does not have any mnemonic effect and cannot be said to be a good predictor of learning after all. This seems to contradict the correlational findings reported in Table 4, but we should not forget that the correlational analyses were based on the post hoc ratings of transparency and imageability, which are very likely to have been affected by participants' test performance. This negative finding underscores the relevance of imageability and the concept of dual coding, which turned out to be more important than transparency for learning.

Long-term Effects of L2 Idiom Learning

In our introduction, we presented different views on the issue of whether delayed performance is boosted or diminished by more difficult learning. De Groot and Keijzer (2000), for instance, found that stimulus material that is more difficult to learn is easier to forget. Schneider et al. (2002), on the other hand, claimed that productive learning, due to its inherent difficulty, resulted in inferior immediate and superior delayed performance on the more difficult productive test. In line with our predictions, direction of learning affected delayed performance, but only on the productive test. By comparing the effect sizes of the two delayed tests, we showed that even 3 weeks after initial learning, productive learners still outperformed receptive learners on the productive

test to a considerable degree and that there was no such difference on the receptive test. These findings bear a great deal of resemblance to the immediate performance findings. The fact that the delayed test scores reflect the same pattern as the immediate test scores shows the stability of the effect of direction of learning over time and is consistent with Schneider et al.'s view that more difficult learning can be expected to result in superior learning over time. This effect, however, could only be observed in the more demanding productive test.

To further explore the issue of stability of retention over time, one could compare immediate test scores with delayed test scores (within subject). This, however, would confound immediate performance with relative amount of forgetting over time. After all, the higher the learners' starting point, the more profound the drop in performance can be. Also, comparing the amount of forgetting (in the period between the immediate and the delayed test) across the experimental conditions is a problematic issue. Again, to use a simplification, the question remains as to whether it is the height from which the level of performance has dropped that should be taken into account or the level reached after the deterioration (in absolute rather than relative terms). This question cannot be properly disentangled with the design we used and is beyond the scope of this article (but see Hulstijn, 2003; Wang & Thomas, 1995; Wang, Thomas, & Ouellette, 1992). Moreover, we only had two measurements over time, which is not enough to estimate what the curve of forgetting might have looked like. On the productive test only, performance scores of participants who learned productively decreased substantially, yet their delayed performance was still better than participants who learned receptively. This is consistent with De Groot and Keijzer (2000), whose data corroborated the claim that stimulus material that is more difficult to learn is easier to forget. Also, it is in accordance with Schneider et al. (2002), as productive learning (the more difficult type) resulted in superior delayed performance on the productive test. The fact that the familiar pattern can be discerned again lends support to the assumption that one and the same mechanism might have been responsible for the similar findings: Initial learning difficulty leads to superior delayed performance on a more demanding test. This issue could be addressed in future research. One possibility for disentangling immediate performance from the range of forgetting over time would be to make sure that participants have learned up to the same level of immediate performance—for example, through repeated learning until everybody, independent of learning direction, reaches 100% immediate performance.

Particularities and Limitations of the Study

Various researchers have studied L2 vocabulary learning in a paired-associate paradigm because "it identifies the component processes in vocabulary learning and it suggests an [*sic*] hypothesis about the locus of difficulty in vocab-

ulary learning and a way to overcome that difficulty" (Schneider et al., 2002, p. 437) or because this procedure is, on the one hand, "an important component of most FL [foreign language] training programs" (De Groot & Keijzer, 2000, p. 2) and, on the other hand, more efficient than other methods, such as the keyword or the picture-naming method (which is claimed to be effective for experienced L2 learners as well). Our reasons for choosing to conduct a PAL experiment were that it affords a favorable opportunity to concentrate on the strength of the associative links created between the L1 and the L2 expressions by means of manipulating translation direction. One advantage that the context-free approach conferred to this experiment was that the danger of incorrect inferencing could be eliminated, but this claim should certainly not be interpreted as a rejection on our part of the utility of context. As with the rest of vocabulary, we believe that with regard to idioms "a variety of contexts will evoke a variety of enriching instantiations" (Nation, 2001, p. 241) and that context can provide additional information about "situations of use and finer aspects of meaning" (Nation, p. 242).

(Received 10 November 2006)

NOTES

1. A study by Titone and Connine (1994) actually reported the descriptive norms of four dimensions (familiarity, compositionality, predictability, and literality) along which idioms can differ. These ratings were provided for 171 English idioms by native English speakers. What we needed, however, were ratings of imageability and transparency, and we collected those in a separate norming study.

2. Due to a technical problem, the dataset of 1 participant (out of 129) is incomplete. In this case, only data concerning that person's condition and performance on the immediate test are available; the rest of the variables could not be recorded properly and have been encoded as missing values for the analyses. For this reason, we can only report on participants' gender for 128 participants.

3. Olkin and Siotani (1964) have shown that the term $z = \mathrm{SQRT}(n) * (r_{ab} - r_{ac})/\mathrm{SQRT}((1 - r_{ab}^2)^2 + (1 - r_{ac}^2)^2 - 2r_{bc}^3 - (2r_{bc} - r_{ab} * r_{ab}) * (1 - r_{ab}^2 - r_{ac}^2 - r_{bc}^2))$ is normally distributed, with a and b referring to the predictors (i.e., transparency and imageability ratings) and c referring to the criterion (i.e., immediate performance). The resulting z-values are shown in Table 4; z-values below -1.96 and above 1.96 are significant at $p < .05$ (two-tailed).

4. We tested whether possible selection effects in the delayed test could have influenced the results. Two questions arise here. First, does the experimental manipulation affect participants' participation in the delayed test? This did not seem to be the case, as response rates did not differ significantly between conditions, $\chi^2(3, 129) = 7.78$, $p = .05$. Second, do people who participate in the delayed test systematically differ from those who do not participate and would such systematic self-selection occur differently in the four experimental conditions? We compared immediate performance scores for three subsamples formed by participants' reaction type in terms of time frame within which they replied; that is, whether participants reacted to the first e-mail ($n = 44$), or whether they reacted only after being reminded once ($n = 14$), or whether they did not react at all ($n = 70$). Immediate performance scores were entered into a 3 (reaction type) $\times$ 4 (experimental condition) ANOVA. Results revealed no general selection effect, $F(2, 117) = 2.53$, $p = .08$. More importantly, there was no interaction between reaction type and condition, $F(6, 117) = 1.00$, $p = .43$. These findings show that even if there might be a slight selection effect such that participants with better immediate performance scores were more likely to participate in the follow-up, this possible selection is independent of the experimental condition; that is, this selection would occur to the same degree in all experimental conditions. Therefore, selection effects cannot account for differences between the experimental conditions.

REFERENCES

Arnaud, P. J. L., & Savignon, S. (1997). Rare words, complex lexical units and the advanced learner. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 157–173). New York: Cambridge University Press.

Battig, W. F. (1979). The flexibility of human memory. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 23–44). Mahwah, NJ: Erlbaum.

Boers, F., Demecheleer, M., & Eyckmans, J. (2004). Etymological elaboration as a strategy for learning idioms. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 53–78). Amsterdam: Benjamins.

Bortfeld, H. (2002). What native and non-native speakers' images for idioms tell as about figurative language. In R. R. Heredia & J. Altarriba (Eds.), *Bilingual sentence processing* (pp. 275–295). Amsterdam: Elsevier.

Bortfeld, H. (2003). Comprehending idioms cross-linguistically. *Experimental Psychology*, *50*, 217–230.

Bortz, J. (1989). *Statistik für Sozialwissenschaftler* [Statistics for social scientists]. Berlin: Springer-Verlag.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Mahwah, NJ: Erlbaum.

Cacciari, C., & Glucksberg, S. (1991). Understanding idiomatic expressions: The contribution of word meanings. In G. B. Simpson (Ed.), *Understanding word and sentence* (pp. 217–240). Amsterdam: Elsevier.

Cacciari, C., & Glucksberg, S. (1995). Imagining idiomatic expressions: Literal or figurative meanings? In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 43–56). Mahwah, NJ: Erlbaum.

Cooper, T. C. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly*, *33*, 233–262.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Crothers, E., & Suppes, P. (1967). *Experiments in second-language learning*. San Diego: Academic Press.

De Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning*, *56*, 463–506.

De Groot, A. M. B., & Hoeks, J. C. J. (1995). The development of bilingual memory: Evidence from word translation by trilinguals. *Language Learning*, *45*, 683–724.

De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*, 1–56.

De Groot, A. M. B., & Poot, R. (1997). Word translation at three levels of proficiency in a second language: The ubiquitous involvement of conceptual memory. *Language Learning*, *47*, 215–264.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*, 559–617.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of Language*, *6*, 22–42.

Geeraerts, D. (1995). Specialization and reinterpretation in idioms. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 57–73). Mahwah, NJ: Erlbaum.

Gibbs, R. W., Jr. (1985). On the process of understanding idioms. *Journal of Psycholinguistic Research*, *14*, 465–472.

Gibbs, R. W., Jr. (1993). Why idioms are not dead metaphors. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 57–77). Mahwah, NJ: Erlbaum.

Gibbs, R. W., Jr., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, *21*, 100–138.

Gibbs, R. W., Jr., Nayak, N. P., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28*, 576–593.

Glucksberg, S. (1993). Idiom meanings and allusional content. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 3–26). Mahwah, NJ: Erlbaum.

Griffin, G., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, *17*, 443–460.

Horowitz, L. M., & Gordon, A. M. (1972). Associative symmetry and second language learning. *Journal of Educational Psychology*, *63*, 287–294.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*, 24–44.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Oxford: Blackwell.

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly*, *20*, 287–304.
Irujo, S. (1993). Steering clear: Avoidance in the production of idioms. *IRAL*, *31*, 205–219.
Jacoby, L. L., & Craik, F. I. M. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 1–21). Mahwah, NJ: Erlbaum.
Jones, G. V. (1985). Deep dyslexia, imageability, and ease of prediction. *Brain and Language*, *24*, 1–19.
Kellerman, E. (1983). Now you see it, now you don't. In S. M. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112–134). Rowley, MA: Newbury House.
Kroll, J. F. (1993). Accessing conceptual representations for words in a second language. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 63–81). Amsterdam: Benjamins.
Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*, 149–174.
Kroll, J. F., & Tokowicz, N. (2005). Models of bilingual representation and processing: Looking back and to the future. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–554). Oxford: Oxford University Press.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
Liontas, J. (2002). Context and idiom understanding in second languages. *EUROSLA Yearbook*, *2*, 155–185.
Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*, 397–403.
Mägiste, E. (1979). The competing language systems of the multilingual: A developmental study of decoding and encoding processes. *Journal of Verbal Learning and Verbal Behavior*, *18*, 79–89.
Mondria, J.-A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 79–100). Amsterdam: Benjamins.
Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). New York: Cambridge University Press.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York: Cambridge University Press.
Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*, 491–538.
Olkin, J., & Siotani, M. (1964). *Asymptotic distribution functions of a correlation matrix* (Rep. No. 6). Stanford, CA: Stanford University Laboratory for Quantitative Research in Education.
Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*, 241–263.
Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, & Winston.
Paivio, A., Yuille, J. C., & Smythe, P. C. (1966). Stimulus and response abstractness, imagery, and meaningfulness, and reported mediators in paired-associate learning. *Canadian Journal of Psychology*, *20*, 362–377.
Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.
Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440.
Schuyten, M. C. (1906). Experimentelles zum Studium der gebräuchlichsten Methoden im fremdsprachlichen Unterricht [Experimental approaches to the study of the most common methods in foreign language teaching]. *Experimentelle Pädagogik*, *3*, 199–210.
Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*, 499–520.
Stoddard, G. D. (1929). An experiment in verbal learning. *Journal of Educational Psychology*, *20*, 452–457.
Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbolic Activity*, *9*, 247–270.
Underwood, B. J., & Schulz, R. W. (1960). *Meaningfulness and verbal learning*. New York: Lippincott.
Wang, A. Y., & Thomas, M. H. (1995). Effect of keywords on long-term retention: Help or hindrance? *Journal of Educational Psychology*, *87*, 468–475.
Wang, A. Y., Thomas, M. H., & Ouellette, J. A. (1992). Keyword mnemonic and retention of second-language vocabulary words. *Journal of Educational Psychology*, *84*, 520–528.
Weinreich, U. (1969). Problems in the analysis of idioms. In J. Puhvel (Ed.), *Substance and structure of language*. Berkeley: University of California Press.

APPENDIX

STIMULUS IDIOMS, THEIR MEANING IN DUTCH, AND DUTCH EQUIVALENTS

English idiom	Meaning in Dutch	Dutch idiom
To get off the hook	Aan een onaangename situatie ontsnappen	De dans ontspringen
To fly off the handle	Plotseling je zelfbeheersing verliezen, je laten gaan	Over de rooie gaan; over zijn toeren raken
To lay something at somebody's door	Iemand de schuld voor een onaangename gebeurtenis of situatie geven	Iets in iemands schoenen schuiven
To paint the town red	Uitbundig feest vieren, stappen	De bloemetjes buiten zetten
To get cold feet (about something)	Angstig worden voor iets omdat je het gevoel hebt dat het verkeerd kan aflopen	Het heen-en-weer krijgen van iets
To hang fire	Een beslissing uitstellen, afwachten	Een plan in de koelkast zetten; iets op de lange baan schuiven
To stick to your guns	Niet van je standpunt afwijken ofschoon andere mensen proberen het jou duidelijk te maken dat je ongelijk hebt	Voet bij stuk houden
To have had your fill of something	Er genoeg van iets hebben gehad, niets meer daarvan willen	Je bent het zat; je hebt het gehad.
To sit on the fence	Het vermijden om de ene of de andere partij in een discussie te steunen	Zich op de vlakte houden
To be in for it	Waarschijnlijk in moeilijkheden gaan zitten vanwege iets dat men gedaan heeft	De bui voelen hangen
To play the field	Liefdesrelaties met meerdere mensen hebben	Van twee walletjes eten; meerdere ijzers in het vuur hebben
To show your hand	Laten zien hoe machtig je bent en hoe je van plan bent te handelen	Je spierballen laten zien; je spierballen laten rollen
To wear your heart on your sleeve	Je gevoelens duidelijk laten zien, ze niet verbergen	Het staat op je gezicht geschreven
To carry the day	De winnaar zijn in een gevecht, debat of wedstrijd	Aan het langste eind trekken
To (manage to) keep a straight face	Ernstig blijven kijken, ofschoon je eigenlijk iets zo grappig vindt dat je in lachen zou kunnen uitbarsten	Je gezicht in de plooi houden
To have gone off the deep end	Zich niet meer normaal gedragen, zich zo gedragen alsof je niet meer goed wijs bent	Van lotje getikt zijn; ze niet allemaal op een rijtje hebben
To put on airs	Je hautain gedragen, alsof je meer of beter bent dan andere mensen	Uit de hoogte doen
To beat someone at their own game	Iemand net zo behandelen als hij anderen behandelt, dezelfde methodes gebruiken, en zelfs meer succes erbij hebben	Iemand een koekje van eigen deeg geven
To be down in the dumps	Zich heel depressief en ongelukkig voelen	Bij de pakken neerzitten
To shoot/fire from the hip	Heel snel en ondoordacht op een situatie reageren of zeggen wat je mening is	Alles zeggen wat je voor de mond komt